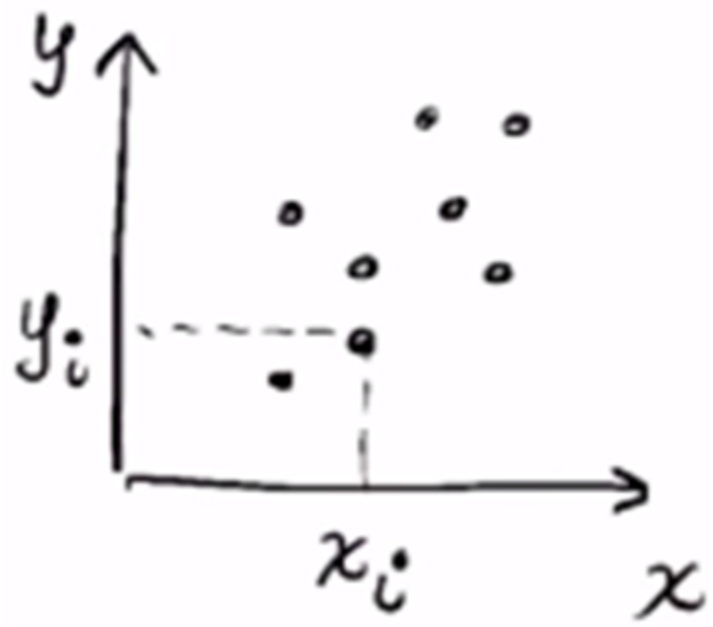
Correlaciones

Recordemos el uso de scatterplot, es decir cuando vimos como una variable se relacionaba con respecto de otra.

Cuando una variable se comporta o tiene la misma dirección de crecimiento o decrecimiento con respecto de otra de una forma bien definido, decimos que están correlacionadas y entonces no tiene sentido incluir ambas variables en un modelo de Machine Learning porque probablemente las 2 variables están aportando la misma información si su correlación es muy alta, por lo cual habría que eliminar alguna de las 2. De aquí es importante hacer un reducción de datos contemplando estos valores de correlación.

Por otro lado decimos que si tenemos variables categóricas y las convierto en numéricas, esto expande mi espacio de atributos numéricos y entonces tendría muchas variables, por lo que tambien puedo buscar si existen correlaciones de una u otra manera, para así reducir el numero de variables, entonces lo que podemos hacer es como comenzar ese proceso de cómo reducir datos hablando del concepto de correlación

Recordando conceptos



$$var(x_1, \ldots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sigma^2$$

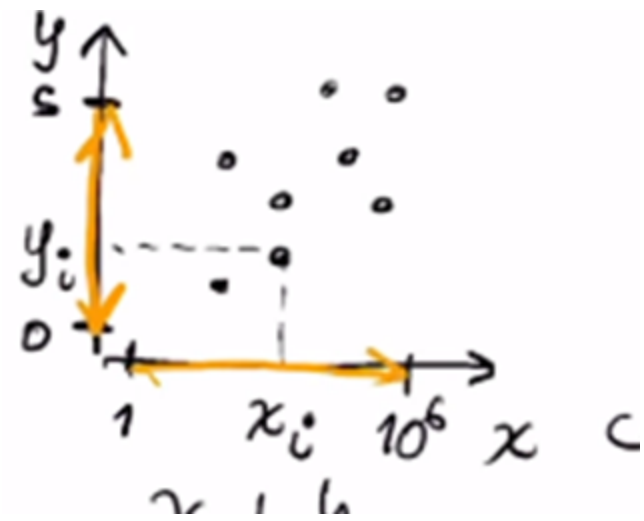
¿Cómo nos ayuda el concepto de varianza a la correlación?

Correlación tiene que ver con medir las desviaciones o variaciones de una variable X en relación a otra variable Y.

Nosotros podríamos estar tomando las variaciones $(X_i - \overline{X})$ de cada elemento X_i y multiplicarlo por las variaciones de la segunda variable con respecto de su promedio. $(X_i - \overline{X})(Y_i - \overline{Y})$. Pensemos esto como tener una tabla con los valores de X vs valores de Y. De esta manera estoy relacionando una cantidad que me entrega resultados de los 2 tipos de variación de las variabas, si esto lo hago de la misma forma que lo hago con la varianza, yo puedo sumar sus efectos individuales para después promediar y obtener un resultado, a esto le llamamos **covarianza** (Porque es la varianza conjunta entre estas 2 variables).

$$cov = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Hay un detalle interesante y es que; cuando estamos trabajando con datos cada uno tiene su componente en x y y puede que estén en distintas unidades o escalas.



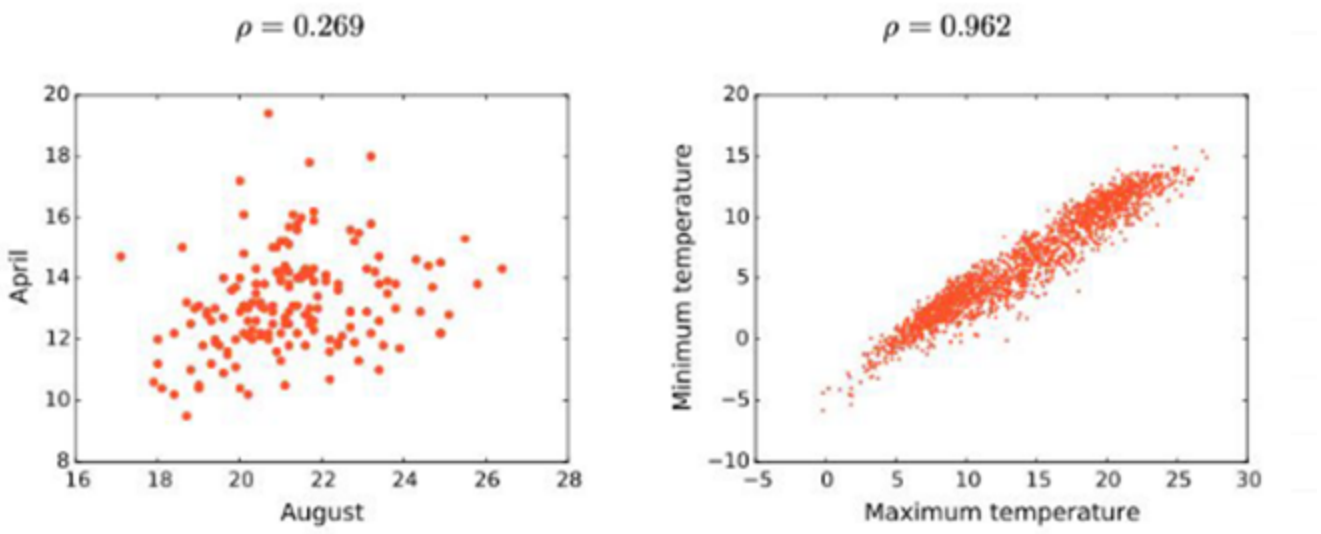
En estos casos los ordenes de magnitud de cada una de las variables son muy distintos, mientras que x va en una escala de $1 - 10^6$ e y va de $0 - 5$, también puede ser que tenga unidades diferentes como longitud y tiempo, etc. Así que una manera mas estandarizada de dar reporte del grado de variación de cada variable es tomar el concepto de **covarianza** y dividir cada una de las variaciones de las variables y dividirlo con la **desviación estándar** σ, para estandarizar las variables a una misma escala; esto conduce a un concepto llamado: "**coeficiente de correlación**"

$$\rho = \frac{cov}{\sigma_x \sigma_y}$$

En donde ρ es la letra griega rho, y denota el coeficiente de correlación:

- cuando ρ La correlación entre las variables es grande
- cuando ρ La correlación entre las variables es pequeña

La siguiente imagen refuerza lo estipulado anteriormente



- Izquierda ρ
- Derecha ρ

En el caso de la izquierda las variaciones en los ejes no muestran una co-relación, por otro lado en el caso de la derecha ambas variables están co-relacionadas clara entre ambas.

> En el caso de **Correlaciones negativas** quiere decir que si una variable aumenta, la otra disminuye.

Entonces decimos que $-1 \leq \rho \leq 1$ indicando:

- $\rho = -1$ Inversamente correlacionado perfectamente.
- $\rho = 1$ Directamente correlacionado perfectamente.
- $\rho \approx 0$ No existe una correlación.

NOTA

Tenemos un parámetro que puede medir la correlación pero de manera simplemente **estadística**, pero cuando uno mira el trasfondo del problema no quiere decir que haya una relación **CAUSA-EFECTO** entre una variable y la otra, es importante tener esto muy claro.

En estadistica hay un lema $\LARGE\{\text{”la correlación no implica causalidad”}\}$.

Entonces no trates de deducir que una variable pueda ser necesariamente causa de la otra, Puede ser o no, el concepto matemático no garantiza eso.

Esto hay que tenerlo en cuenta para seleccionar de manera adecuada si hay ocupar alguna variable en especifico para el modelo que se desee implementar.

Extras

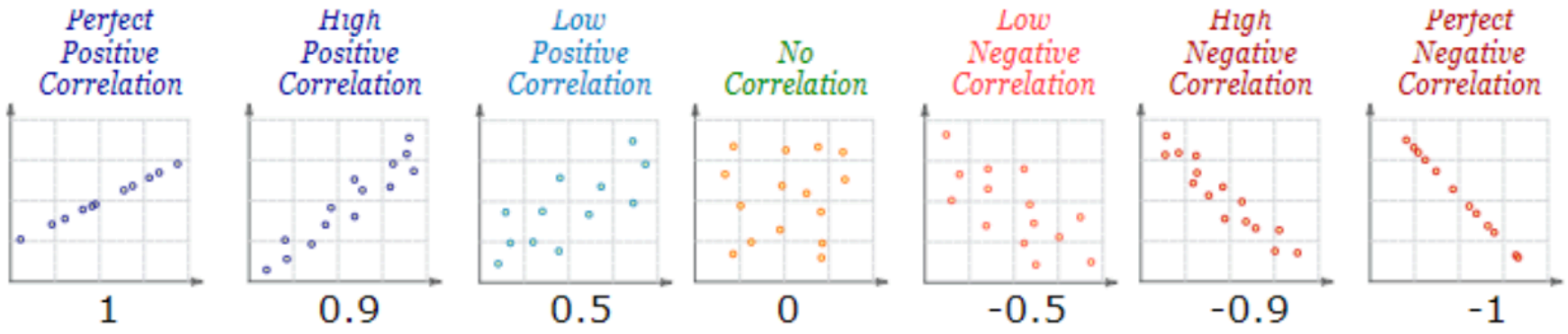
Aporte de David Sánchez lombana

Correlaciones
¿Qué es la correlación?
La correlación es una medida estadística que expresa hasta qué punto dos variables están relacionadas linealmente (esto es, cambian conjuntamente a una tasa constante).

¿Qué es la covarianza?
Es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.

¿Qué es el coeficiente de correlación? El coeficiente de correlación es la medida específica que cuantifica la intensidad de la relación lineal entre dos variables en un análisis de correlación.

Correlación gráficamente:



- Co-relación y Regresión lineal